Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators

Year: 2023 | Citations: 711 | Authors: Levon Khachatryan, A. Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang W

Abstract

Recent text-to-video generation approaches rely on computationally heavy training and require large-scale video datasets. In this paper, we introduce a new task, zero-shot text-to-video generation, and propose a low-cost approach (without any training or optimization) by leveraging the power of existing text-to-image synthesis methods (e.g. Stable Diffusion), making them suitable for the video domain. Our key modifications include (i) enriching the latent codes of the generated frames with motion dynamics to keep the global scene and the background time consistent; and (ii) reprogramming frame-level self-attention using a new cross-frame attention of each frame on the first frame, to preserve the context, appearance, and identity of the foreground object. Experiments show that this leads to low overhead, yet high-quality and remarkably consistent video generation. Moreover, our approach is not limited to text-to-video synthesis but is also applicable to other tasks such as conditional and content-specialized video generation, and Video Instruct-Pix2Pix, i.e., instruction-guided video editing. As experiments show, our method performs comparably or sometimes better than recent approaches, despite not being trained on additional video data. Our code is publicly available at: https://github.com/Picsart-AI-Research/Text2Video-Zero.